A wearable vision-based system for detecting hand-object interactions in individuals with cervical spinal cord injury: First results in the home environment

Andrea Bandini, *Member, IEEE,* Mehdy Dousty, and José Zariffa, *Senior Member, IEEE*

***Abstract*— Cervical spinal cord injury (cSCI) causes the paralysis of upper and lower limbs and trunk, significantly reducing quality of life and community participation of the affected individuals. The functional use of the upper limbs is the top recovery priority of people with cSCI and wearable vision-based systems have recently been proposed to extract objective outcome measures that reflect hand function in a natural context. However, previous studies were conducted in a controlled environment and may not be indicative of the actual hand use of people with cSCI living in the community. Thus, we propose a deep learning algorithm for automatically detecting hand-object interactions in egocentric videos recorded by participants with cSCI during their daily activities at home. The proposed approach is able to detect hand-object interactions with good accuracy (F1-score up to 0.82), demonstrating the feasibility of this system in uncontrolled situations (e.g., unscripted activities and variable illumination). This result paves the way for the development of an automated tool for measuring hand function in people with cSCI living in the community.**

***Clinical relevance*— The accurate detection of hand-object interactions in people with cSCI will allow extracting outcome measures of upper limb function that can be used for planning interventions and tracking the rehabilitation progress remotely.**

I. INTRODUCTION

The functional use of the upper limbs is the top recovery priority in people with cervical spinal cord injury (cSCI) [1]. However, two main factors may prevent the optimal recovery of upper limb functions: 1) patients are often discharged too early, when they may still experience large improvements in their functional ability [2]; 2) particularly in North America, the long distances between patients' homes and rehabilitation centers make it difficult to accurately track the recovery of individuals with cSCI [3]. These factors hinder the rehabilitation process and do not allow planning optimal treatment strategies for improving upper limb functions when people return to the community. Hence, there is an urgent need for novel technologies able to produce accurate outcome measures of upper limb function in individuals with cSCI. The ideal solution should be able to track the rehabilitation progress remotely, producing measures that can be used by clinicians for planning interventions. It should also be able to measure how improvements in the upper limb motor functions actually translate into an increased usage in daily life (i.e., measuring performance in addition to capacity, in the terminology of the International Classification of Functioning, Disability and Health [4]).

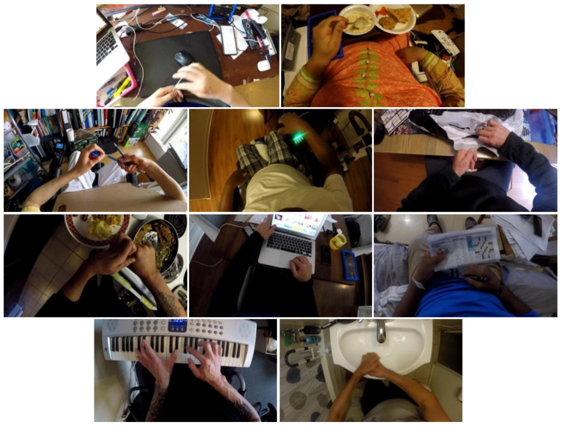

Fig. 1. Example of egocentric frames obtained from participants with cSCI while performing unscripted activities in their homes.

Wearable sensors such as accelerometers, magnetometers, and inertial measurement units (IMUs) have been used to fill this gap. For example, clinical assessment scales (e.g., the functional ability scale) were estimated in stroke survivors by using accelerometer data from upper limb movements [5], [6], whereas IMUs were used in patients with acute SCI for monitoring their upper limb and wheeling activities [7], [8]. Although well-established and easy-to-implement for long and continuous recordings, these systems allow extracting only global kinematic information of the upper limbs, with no details regarding hand manipulations or finger movements. Other approaches either combined multiple sensors (e.g., wrist-worn accelerometer and finger magnetometer [9]) or used sensorized gloves [10] to capture the hand and finger motion in higher detail. However, their usage may become inconvenient in people whose hand function and sensation is already impaired due to cSCI.

Wearable cameras (i.e., cameras mounted on the head)

*This work was supported in part by the Craig H. Neilsen Foundation (542675).

[1]A. Bandini is with KITE – Toronto Rehab – University Health Network, Toronto, ON, CA. andrea.bandini@uhn.ca

[2]M. Dousty is with KITE – Toronto Rehab – University Health Network, and the Institute of Biomaterials and Biomedical Engineering (IBBME), University of Toronto, Toronto, ON, CA mehdy.dousty@mail.utoronto.ca

[2]J. Zariffa is with KITE – Toronto Rehab – University Health Network; IBBME, University of Toronto; the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto; and the Rehabilitation Sciences Institute, University of Toronto, Toronto, ON, CA. jose.zariffa@utoronto.ca

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

and computer vision have emerged as potential candidates for measuring the hand use at home [11], [12]. From the egocentric point of view (POV), it is possible to observe the hands from a perspective that minimizes the occlusions and focuses on the hands and manipulated objects [11] (Figure 1). This field of research has thrived over the past ten years, thanks to the availability of action cameras and large annotated datasets. Several approaches have been proposed to localize the hands within the egocentric video frames, understand the type of hand grasps and gestures, predict the actions and activities involving hand manipulations, as well as develop applications for human-machine interaction [11]. Recent findings [13], [14] showed that the use of hand and object cues automatically extracted from egocentric videos (e.g., color, motion, and edges features) allowed detecting with good accuracy the presence of hand-object interactions. This approach can be used to extract novel metrics that, once validated with clinical gold standards, will constitute clinically valid outcome measures for upper limb function in cSCI. However, the approaches proposed so far [13], [14] were tested only in a home simulation laboratory that, although realistic, did not present the common issues of natural environments such as high variability of background, illumination, objects, and activities.

Building upon the previous results [13], [14], we exploit the recent advancements in computer vision – in particular the availability of accurate object recognition convolutional neural networks (CNNs) – to develop a deep learning-based approach for detecting hand-object interactions in individuals with cSCI living in the community. For the first time we demonstrate its feasibility in individuals with cSCI who recorded unscripted activities in their homes, without the supervision of researchers.

II. MATERIALS AND METHODS

A. Data Collection

Ten individuals with cSCI – 8 male and 2 female – were recruited for this study. Their age ranged between 42 and 63 years old (51.0 ± 8.4 years). Seven participants had a traumatic injury, whereas for 3 participants the etiology was non-traumatic. The number of months from the injury ranged between 18 and 264 (103.6 ± 90.3 months). Before the experiments, participants were assessed using the International Standards for Neurological Classification of SCI (ISNCSCI) [15], the Spinal Cord Independence Measure III (SCIM) [16], and the Graded Redefined Assessment of Strength, Sensibility and Prehension (GRASSP) [17] (Table I).

TABLE I
CLINICAL INFORMATION OF THE PARTICIPANTS RECRUITED IN THIS STUDY.

Level of Injury	C4 – C8
ASIA Impairment Scale (AIS)	A – D
Upper Extremity Motor Score (UEMS)	Right: 19.3 ± 6.0 Left: 19.1 ± 6.3
GRASSP	Right: 82.8 ± 29.9 Left: 86.6 ± 31.0
SCIM	70.9 ± 29.3

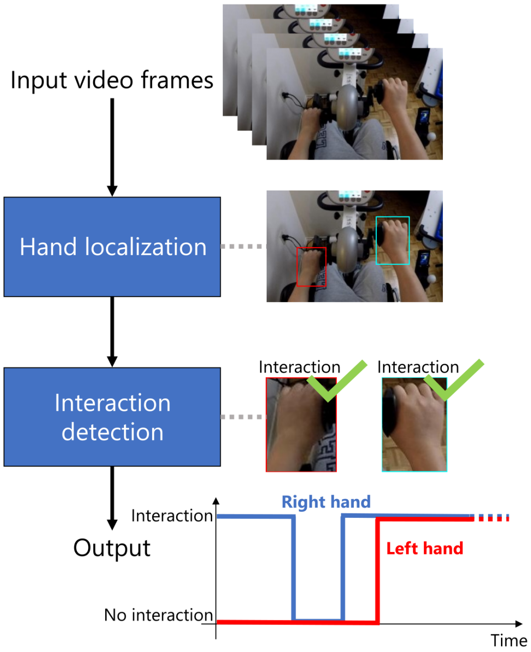

Fig. 2. A diagram of the processing steps for detecting hand-object interactions from egocentric videos.

Participants recorded activities that involve the use of hands in their homes using a head-mounted camera (GoPro$^{®}$ Hero 5 Black – Figure 1). They were asked to record 3 videos of at least 1.5 hours each (over two weeks), during activities of their normal daily routine. Examples of activities collected for this study included: feeding; grooming and personal hygiene (e.g., brushing teeth, washing hands); functional mobility; home establishment and management (e.g., doing the laundry, cleaning); meal preparation; and leisure activities (Figure 1). Videos were recorded at 1080p resolution and 30 frames per second. The study was approved by the Research Ethics Boards at UHN – Toronto Rehabilitation Institute. All participants and any household members appearing in the videos signed informed consent according to the requirements of the Declaration of Helsinki.

B. Video Processing – Interaction Detection Evaluation

The hand-object interaction detection pipeline is composed of two processing steps: 1) the hand localization – to detect the participant's hands within each frame and focus the processing on these regions of interest (ROIs); and 2) the interaction detection – to predict whether the detected hands are interacting with objects (Figure 2).

First, we conducted a test to evaluate the performance of the interaction detection module alone, using the ground truth bounding boxes (labelled by a trained annotator) as hand localization result. Specifically, two approaches were implemented and compared: a multistage image processing approach (baseline) and an end-to-end deep-learning approach (Hand-Object Interaction Detection Network – HOID-NET).

The baseline approach was previously proposed in [14] and exploits color, optical flow, and histogram of oriented gradient (HOG) features extracted from the hand region, the region next to the hand (i.e., the manipulated object), and the background. A random forest classifier was used to detect the presence of interactions between hands and objects. Our newly proposed approach – HOID-NET – was implemented by using transfer learning on the pre-trained MobileNet v1 [18], a compact yet powerful object recognition CNN. We chose this architecture for its good recognition performance with low resource devices, which may allow us in the future to implement the whole processing pipeline on a portable embedded system. Specifically, we replaced the output layer with 3 fully connected layers (with 1024, 1024, and 512 neurons, respectively), followed by a softmax layer with 2 neurons to produce the binary predictions (*Interaction* vs. *No interaction*). Starting from the pre-trained ImageNet weights, we trained the network on our dataset for 20 epochs using ADAM optimizer. The learning rate was set at 10^{-3} and halved every 3 epochs, with batch size equal to 16.

Results obtained with the two approaches were compared in terms of accuracy, precision, recall and F1-score obtained on the test set. These measures were calculated on the whole test set, without distinction of the different activities recorded by the participants.

1) Dataset: The dataset used for the interaction detection evaluation was composed of 72,203 frames. All frames were resized to 720 × 405 pixels. For the HOID-NET, we considered the frames cropped using the coordinates of the hand bounding boxes, for a total of 103,741 hand samples. These images were resized to 224 × 224 pixels to match the CNN input layer. The dataset was split as follows: training set – 46,463 frames from 3 participants (plus 17 participants recorded in a home simulation environment – ANS SCI dataset [14]), corresponding to 64,979 hand instances; validation set – 13,999 frames from 3 participants, corresponding to 24,018 hand images; test set – 11,741 frames from 3 participants, corresponding to 14,744 hand images.

C. Video Processing – Fully Automated Pipeline

After the best interaction detection approach had been identified, we tested the fully automated pipeline on an additional participant. The localization of hands was performed using the approach proposed in [19], which exploited the *You Only Look Once* (YOLOv2) object detector [20] trained on the ANS SCI dataset [14], on which it yielded excellent detection performance (F1-score = 0.88). Moreover, the binary output of the interaction detection pipeline was further processed using a moving average filter, in order to remove short and isolated sequences with *Interactions* or *No Interactions* predictions.

We tested the full pipeline on data from one participant left out from the experiments reported in Sec. II-B. Specifically, we considered 31,368 frames, corresponding to 62,707 hand instances.

III. RESULTS

The performance of the two interaction detection approaches are reported in Table II. HOID-NET produced better detection results than the baseline method, suggesting the feasibility of using a CNN-based approach to infer the presence of hand-object interactions from egocentric videos recorded in an uncontrolled environment.

TABLE II
INTERACTION DETECTION PERFORMANCE ON THE TEST SET, WHEN USING THE GROUND TRUTH BOUNDING BOX TO LOCALIZE HANDS (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD).

Methods	Accuracy	Precision	Recall	F1-score
Baseline [14]	0.6496	0.7533	0.7623	0.7578
HOID-NET	**0.7257**	**0.7831**	**0.8551**	**0.8175**

Before implementing the fully-automated pipeline we also evaluated the hand localization performance of YOLOv2 on data recorded at home (32,013 frames manually labelled from 9 individuals with cSCI). The F1-score was 0.87, comparable to the one obtained on the ANS SCI dataset [19]. This result suggests that this algorithm is robust in localizing the hands even in uncontrolled situations.

The interaction detection performance obtained with the fully-automated pipeline implemented on the test participant is reported in Table III.

TABLE III
INTERACTION DETECTION PERFORMANCE ON A TEST SUBJECT, WHEN USING THE FULLY AUTOMATED PIPELINE (HAND LOCALIZATION = YOLOV2; INTERACTION DETECTION = HOID-NET).

Accuracy	Precision	Recall	F1-score
0.6745	0.7265	0.7907	0.7573

IV. DISCUSSION

In this paper we proposed HOID-NET, a CNN-based approach to detect hand-object interactions from egocentric videos recorded at home by individuals with cSCI. HOID-NET produced better results than the multi-stage image processing approach, by using only the color information from the RGB frames. This indicates that the hand ROIs and their appearance carry significant information for detecting whether the hands are interacting with objects, allowing to simplify considerably the video-processing steps with respect to [14]. However, as pointed out by recent studies on action recognition [21], [22], the inclusion of the temporal information allows improving the recognition performance. Since our task can be seen as the binary simplification of action recognition (i.e., we only detect when hand actions occur), we believe that the use of 3DCNNs and recurrent neural networks will further boost the detection performance.

The fully-automated pipeline showed good results, with F1-score around 0.75. The decrease of performance from the case with ground truth bounding boxes is due to the error propagation caused by non-perfect hand localization. However, this result is in line with those previously obtained with the baseline approach on the ANS SCI dataset (F1-score between 0.73 and 0.74 [14]) where the illumination conditions and activities were the same for all participants. Thus, it is reasonable to expect that the expansion of the dataset will help improving the perfomance of the system. Interestingly, we noticed that both tests (i.e., with ground truth hand detection – Table II, and YOLOv2 hand detection – Table III) yielded higher recall than precision. This indicates that the system is better at recognizing interactions than no interactions. An explanation for this result can be attributed to presence of non-standard hand postures and gestures (e.g., presence of spasticity), which make it difficult to recognize when the hands are not interacting, leading to the misclassification of some frames.

The dataset expansion will also allow taking into account the large inter-subject variability that exists in people with cSCI. Specifically, the type of gestures and hand postures with which individuals with cSCI interact with objects greatly depend on the level and severity of the injury, as well as on the strategies learned during the rehabilitation process. Thus, the inclusion of a larger and heterogeneous sample of participants will certainly have beneficial effects on the performance and robustness of this system. This is an important factor that needs to be addressed in view of developing and automated tool for monitoring upper limb functions in people with cSCI living in the community.

Besides expanding the dataset and improving the performance of the algorithm through different deep-learning architectures, the future steps will focus on the clinical validation of this system. Some simple measures, such as the number of interactions per unit of time or the duration of interactions proposed in [14] will be extracted and correlated with gold-standard clinical scores (e.g., GRASSP, UEMS, etc), in order to produce validated metrics for monitoring the upper limb functions remotely.

V. CONCLUSIONS

In this work, we demonstrated the feasibility of using an egocentric video-based approach for detecting hand-object interactions in individuals with cSCI living in the community. Testing this technology at home is an important step towards developing an automated tool for monitoring hand function in people with upper limb impairment due to cSCI. The expansion of the dataset and the implementation of temporal models will allow improving the performance of the video-processing algorithms, with the ultimate goal of enabling continuous and remote assessment of upper limbs for individuals with cSCI living in the community.

ACKNOWLEDGMENT

The authors would like to thank Angy Higgy, John Salama, and Chantel Campbell for their valuable assistance in data annotation. The authors would also like to thank all the participants and their families involved in the study.

REFERENCES

[1] G. J. Snoek, M. J. Ijzerman, H. J. Hermens, D. Maxwell, and F. Biering-Sorensen, "Survey of the needs of patients with spinal cord injury: impact and priority for improvement in hand function in tetraplegics," *Spinal Cord*, vol. 42, no. 9, pp. 526–532, 2004.

[2] A. S. Burns and J. F. Ditunno, "Establishing Prognosis and Maximizing Functional Outcomes After Spinal Cord Injury," *Spine*, vol. 26, no. Supplement, 2001.

[3] C. Craven *et al.*, "Rehabilitation Environmental Scan Atlas: Capturing Capacity in Canadian SCI Rehabilitation," Vancouver, BC, Canada: Rick Hansen Institute, 2012.

[4] R. J. Marino, "Domains of outcomes in spinal cord injury for clinical trials to improve neurological function," *The Journal of Rehabilitation Research and Development*, vol. 44, no. 1, p. 113, 2007.

[5] S. Patel, *et al.*, "A Novel Approach to Monitor Rehabilitation Outcomes in Stroke Survivors Using Wearable Technology," *Proceedings of the IEEE*, vol. 98, no. 3, pp. 450–461, 2010.

[6] M. Noorkõiv, H. Rodgers, and C. I. Price, "Accelerometer measurement of upper extremity movement after stroke: a systematic review of clinical studies," *Journal of NeuroEngineering and Rehabilitation*, vol. 11, no. 1, p. 144, 2014.

[7] M. Brogioli, *et al.*, "Monitoring Upper Limb Recovery after Cervical Spinal Cord Injury: Insights beyond Assessment Scores," *Frontiers in Neurology*, vol. 7, 2016.

[8] W. L. Popp, *et al.*, "A novel algorithm for detecting active propulsion in wheelchair users following spinal cord injury," *Medical Engineering & Physics*, vol. 38, no. 3, pp. 267–274, 2016.

[9] N. Friedman, J. B. Rowe, D. J. Reinkensmeyer, and M. Bachman, "The Manumeter: A Wearable Device for Monitoring Daily Use of the Wrist and Fingers," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1804–1812, 2014.

[10] N. P. Oess, J. Wanek, and A. Curt, "Design and evaluation of a low-cost instrumented glove for hand function assessment," *Journal of NeuroEngineering and Rehabilitation*, vol. 9, no. 1, p. 2, 2012.

[11] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[12] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no, 5, pp.744-760, 2015.

[13] J. Likitlersuang and J. Zariffa, "Interaction Detection in Egocentric Video: Toward a Novel Outcome Measure for Upper Extremity Function," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 561–569, 2018.

[14] J. Likitlersuang, E. R. Sumitro, T. Cao, R. J. Visée, S. Kalsi-Ryan, and J. Zariffa, "Egocentric video: a new tool for capturing hand use of individuals with spinal cord injury at home," *Journal of NeuroEngineering and Rehabilitation*, vol. 16, no. 1, May 2019.

[15] S. C. Kirshblum *et al.*, "International standards for neurological classification of spinal cord injury (revised 2011)," *The Journal of Spinal Cord Medicine*, vol. 34, no. 6, pp. 535-546, November, 2011.

[16] A. Catz *et al.*, "A multicenter international study on the Spinal Cord Independence Measure, version III: Rasch psychometric validation," *Spinal Cord*, vol. 45, no. 4, pp. 275-291, April, 2007.

[17] S. Kalsi-Ryan *et al.*, "The Graded Redefined Assessment of Strength Sensibility and Prehension: reliability and validity," *Journal of Neurotrauma*, vol. 29, no. 5, pp. 905-914, March 20, 2012.

[18] A.G. Howard, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint*, arXiv:1704.04861, 2017.

[19] R.J. Visée, J. Likitlersuang, and J. Zariffa, "An effective and efficient method for detecting hands in egocentric videos for rehabilitation applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28 (3), pp.748-755, 2020.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016.

[21] S. Urabe, K. Inoue, and M. Yoshioka, "Cooking activities recognition in egocentric videos using combining 2DCNN and 3DCNN," *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management - CEA/MADiMa* 18, 2018.

[22] Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou, "Multi-Stream Deep Neural Networks for RGB-D Egocentric Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3001–3015, 2019.